多类型分类器融合的文本分类方法研究 *

李惠富，陆　光†

(东北林业大学 信息与计算机工程学院, 哈尔滨 150040)

摘　要： 传统的文本分类方法大多数使用单一的分类器，而不同的分类器对分类任务的侧重点不同，就使得单一的分类方法有一定的局限性，同时每个特征提取方法对特征词的考虑角度不同。针对以上问题，提出了多类型分类器融合的文本分类方法。该模型使用了 word2vec、主成分分析、潜在语义索引以及 TFIDF 特征提取方法作为多类型分类器融合的特征提取方法。并在多类型分类器加权投票方法中忽略了类别信息的问题，提出了类别加权的分类器权重计算方法。通过实验结果表明，多类型分类器融合方法在二元语料库、多元语料库以及特定语料库上都取得了很好的性能，类别加权的分类器权重计算方法比多类型分类器融合方法在分类性能方面提高了 1.19%。

关键词： 文本分类；分类器融合；主成分分析；潜在语义索引

中图分类号： TP391　　doi: 10.3969/j.issn.1001-3695.2017.09.0908

Research on text classification method of multi-class classifier fusion

Li Huifu, Lu Guang†

(*College of Computer & Engineering*, *Northeast Forestry University*, *Harbin* 150040, *China*)

Abstract: Most of the traditional text classification methods use a single classifier, and different classifiers have different emphasis on classification tasks, which makes the single classification method have some limitations. At the same time, each feature extraction method has different angles of considering the feature words. Aiming at the above problems, this paper proposes a text classification method based on multi type classifier fusion, which combines Word2vec, Principal Component Analysis, Latent Semantic Indexing and TFIDF feature extraction as feature extraction methods for the multi type classifier fusion. The weighted voting method of multi type classifier ignores the category information. This paper proposes a weighted classifier weight calculation method. The experimental results show that the multi classifier fusion method has achieved good performance both in two dimensional, multiple corpora and corpus specific corpus, the classification weighting method of classifier weighting improves the classification performance by 1.19% compared with the multi type classifier fusion method.

Key Words: text classification; classifier fusion; principal component analysis; potential semantic index

0　引言

随着互联网的逐渐成熟和微博等社交网络的发展，以信息技术的革命极大了改变了人们的生活方式，越来越多的用户通过网络发布信息和评价实时信息，这些信息的主题类别包括色情、邪教、毒品在内的各种有害信息[1]。因此，如何有效的管理信息对互联网的发展是有重要意义的。

机器学习中的文本分类方法是处理和管理文档数据的关键技术[2]。研究者对文本分类方法进行了广泛的研究。例如，利用 K 最近邻（KNN）方法的简单、无参数等优点对文本垃圾短信分类[3]。Goudjil 等人[4]通过使用 SVM 分类器提供的后验概率来选择样本，利用选择的样本进行分类。在文献[5]中，使用训练数据中深度计算特征加权频率来估计贝叶斯的条件概率，提高了分类的性能。

上述研究方法都取得了很好的分类效果。但是，KNN 方法中的 K 值是人工设置的，具有很大的客观性。SVM 中如何确定高维空间的核函数是目前难点之一。贝叶斯分类中，特征计算时假设特征之间相互独立，而现实中的特征之间是有联系的。上述方法都是采用单一的分类器对文本进行分类，而文本数据涉及的领域非常广，这使得单一分类器不能很好的覆盖更多的领域。因此，本文引进了多分类器融合的文本分类方法。

文本分类是将文本内容相似的文档分配到一个或多个预定义的类别中，而特征提取方法在提高分类器的性能方面有重要的作用[6]。如，文献[7]中，利用动能定理和 TFIDF 特征提取方法来解决微博主题检测问题。文献[8]中，利用 word2vec 作为自动特征提取工具，然后利用句子向量来完成分类。Santosh 等

基金项目：黑龙江省自然科学基金资助项目（F201201）

作者简介：李惠富（1992-），男，黑龙江讷河人，硕士研究生，主要研究方向为文本挖掘；陆光（1963 年-），男（通信作者），副教授，博士，主要研究方向为电子商务与系统开发（lg603@msn.com）.

人[9]利于特征本体树和 LDA 作为在线产品评论文本的特征提取方法，能更好的识别意见词。Uysal 等人[10]使用遗传算法和 LSI 相结合的方法能更好的获取文档的特征向量，能更好的完成分类任务。

上述特征提取方法研究中，都在各自的分类任务中取得了良好的效果。但是，TFIDF 方法只考虑了词语的的统计指标，没有考虑到特征词的语义知识。LSI 方法只考虑了特征词之间的语义关系。word2vec 没有考虑特征词的统计特征。LDA 没有将类别信息加入到主题模型中。上述方法都是单一的特征提取方法并且每个特征提取方法的侧重特征词不同，所以上述方法有一定的局限性。因此，为了能更好的表示文本的特征，本文使用融合的特征提取方法。

综上所述，针对文本分类器和特征提取方法比较单一的问题，本文提出了多类型分类器融合的文本分类方法。并针对分类器加权投票决策方法中没有考虑到分类器对每个类别的贡献度不同的问题，提出了一个类别加权的分类器权重计算方法。

1 多类型分类器融合的文本分类

多类型分类融合的方法是通过融合不同的特征提取方法使得特征空间向量中的特征词更加丰富，通过融合多特征提取方法使文本的表现形式更加丰富，然后通过使用分类器进行文本分类。本文多类型分类融合的方法包含了以下特征提取方法：word2vec；TF-IDF(词频-逆文档频率)；主题模型(latent Dirichlet allocation，LDA)；潜在语义索引(latent semantic indexing，LSI)；

1.1 特征提取方法

1.1.1 word2vec 词向量

在 2013 年，Mikolov 等人提出了一种 word2vec 的开源软件[11]。Word2vec(word to vector)方法通过神经网络方法将单词转换为词向量。在训练词向量的过程中，首先提取出训练数据集中的词语生成词语表，通过使用 CBOW 或者 Skip-Gram 模型来得出每个词语的词向量，模型示意图如图 1 所示。

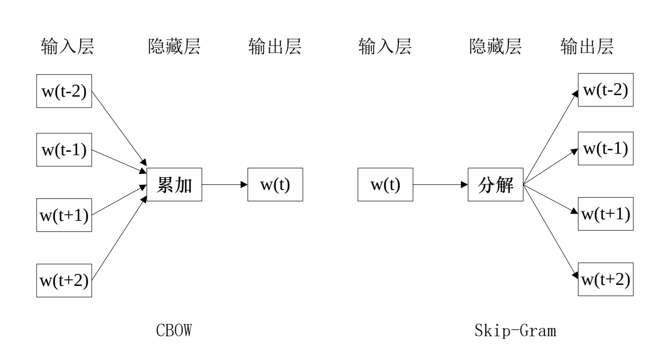


图 1 CBOW 和 Skip-Gram 模型

在图 1 中显示，CBOW 和 Skip-Gram 模型是一个反向的过程。CBOW 模型是利用待预测词前后各 t 个词去预测当前词，而 Skip-Gram 模型是利用当前预测词去预测前后各 t 个词。由于本文使用了 CBOW 模型，所以本文详细介绍了 CBOW 模型。

CBOW 模型是一种利用上下文的信息来预测当前词语出现概率的模型。该模型是一个三层的神经网络，分别为输入层、隐藏层和输出层。输入层是输入词向量，

该词向量为随机值，通过训练数据不断更新词向量。隐藏层是对词向量进行累加。输出层输出词语的概率。

1.1.2 TF-IDF

TF-IDF 是经典的特征权值计算方法，TF-IDF 由 *TF*(词频)和 *IDF*(逆文档频率)构成的，公式如下：

$$tfidf\left(w\right)=tf\left(w\right)\times idf\left(w\right) \tag{1}$$

其中：*tf(w)*为单词 *w* 在文本中出现的次数，*idf(w)*为单词 *w* 的逆文档频率，*idf(w)*的计算方法如式（2）所示。

$$idf\left(w\right)=\log\frac{A}{B\left(w\right)} \tag{2}$$

其中：*A* 代表训练集中文本总的数量，*B(w)*代表包含该词语 *w* 的文件数量。

1.1.3 LDA

LDA 是一种主题模型，是词-文档-主题的三层贝叶斯模型。主题模型通过训练集训练得出主题的 Dirichlet 分布和主题与词之间的多项式分布函数。该方法首先确定一个主题，然后在主题中选择一个单词直到遍历所有的单词，LDA 模型如图 2 所示。

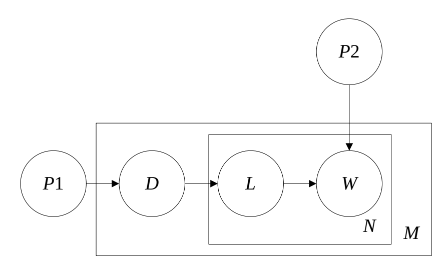


图 2 LDA 模型

图中，*W* 代表文本中的单词；*N* 代表文本中的单词；*M* 代表文本的数量；L 代表参数值 *D* 的多项分布；*P2* 代表 Dirichlet 分布的先验参数，表示词 *W* 的概率；*D* 代表以 *P1* 参数的 Dirichlet 主题分布；*P1* 是 *D* 的参数；

1.1.4 LSI

LSI 是一种无监督的数据挖掘技术，针对一词多义等语义问题有很好的效果。在潜在语义索引方法中，使用奇异值分解方法分解特征向量空间来达到降维的目的，算法模型如图 3 所示。

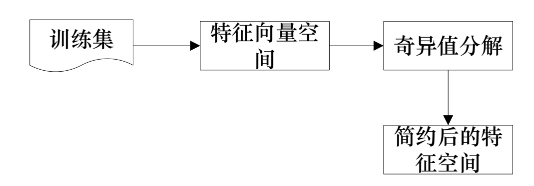


图 3 潜在语义索引模型示意图

1.2 多类型分类器融合

通过 1.1 节的特征提取方法计算使得产生了 4 组不同的特

征向量空间。第一个是由 CBOW 方法产生的 word2vec 的向量空间，第二个是利用 TFIDF 产生的向量空间，第三个是 LSI 产生的语义向量空间，最后一个是利用主题模型产生的 LDA 向量空间。多类型分类器融合的方法是利用特征提取方法产生向量空间之间的互补性。多类型分类器方法模型如图 4 所示，三角形中的数字为分类器的权重。

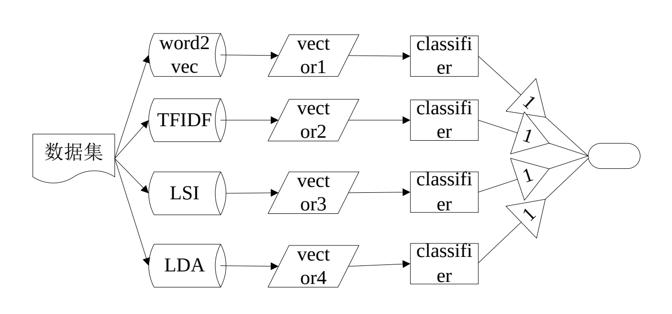


图 4　多类型分类器融合示意图

1.3　类别加权多类型分类器融合

多分类器融合可以利用不同分类器来完成不同的任务，从而避免考虑不全面的问题，不同分类器对同一样本的分类能力不同，因此，每个分类器对每个样本都有不同的贡献能力，分类器加权投票方法是作为组合分类器投票决策的方法之一[12]。分类性能(训练完的样本对训练样本的正确识别率)作为分类器加权投票方法有如下优点：不同分类器对同一样本的识别率是不同的，当组合分类器进行分类决策时，分类结果倾向与分类性能好的分类器，使得决策的性能最好。因此，本文使用分类性能作为分类器的权重。分类性能公式如式（3）（4）所示。

$$\varepsilon = \frac{errorNum}{textNum} \tag{3}$$

$$\alpha = \ln\left(\frac{1-\varepsilon}{\varepsilon}\right) \tag{4}$$

其中：*errorNum* 为分类器未正确分类的样本数目；*textNum* 为样本数据集中样本的总数量；α 为分类器对数据集的权重。

图 5 为传统的二元分类样本示意图，图中一共有 100 个数据点，每个类别有 40 个训练数据点，每个类别有 10 个测试数据点，其中·代表训练数据类别 1，×代表训练数据类别 2，Δ代表测试数据类别 2 中的测试数据，⊲ 为测试数据类别 1 中的测试数据。本部分用 KNN 和多项式贝叶斯作为分类算法，分类性能作为分类器权重。通过 KNN 方法得到了 ε 为 0.0375，α 等于 3.2452，多项式贝叶斯方法得到了 ε 为 0.9625，α 等于 2.5123。因此可知，KNN 分类器权重为 3.2452，多项式贝叶斯分类器权重是 2.5123。KNN 方法的错误样本个数为 12 个，多项式贝叶斯方法的错误样本个数为 10 个，当测试样本点输入组合分类中时，根据投票原则可得该组合方法的错误测试样本个数为 12 个，即是 KNN 方法的错误率。这种方法只是将整个分类器的分类性能作为分类器的权重，忽略了类别对分类器的影响。因此，本文提出了类别加权的分类器权重计算方法。

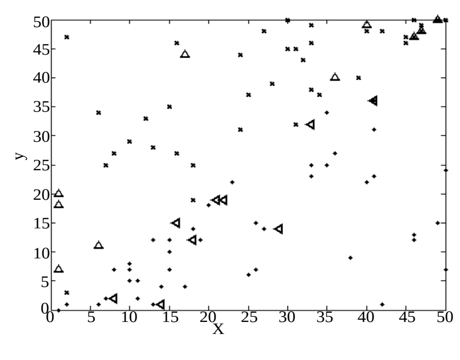


图 5　二元分类样本示意图

类别加权分类器是考虑类别信息对分类器权重的影响。在上述组合分类器中，KNN 分类性能中包含了 1 个负类样本和 2 个正类样本。多项式贝叶斯分类性能中包含了 6 个负类样本。通过上述分析得到，多项式贝叶斯分类器对正类样本有好的识别率，所以想增加多项式贝叶斯正类样本的权重，减少负类样本的分类器权重。因此，本文以不同样本的类别赋予不同的分类器权重。类别加权分类器公式如式(5)所示。通过式(5)计算可以得出，KNN 的正类的 ε 为 0.05，负类的 ε 为 0.025，正类的 α 是 2.9444，负类的 α 是 3.6636。多项式贝叶斯的正类 ε 为 0，负类的 ε 为 0.15，正类的 α 是 10(表示无穷大)，负类的 α 是 1.7346。KNN 和多项式贝叶斯组合方法当测试样本为正类的时候，多项式贝叶斯起到很好的作用，当测试样本为负样本时，KNN 方法对组合分类器的影响较大，根据投票原则得出了该组合方法的错误样本个数为 9 个，相比以前的分类器权重计算方法分类效果得到了提升。

$$w_{li} = \begin{cases} \alpha, x_i \in L_l \\ 0, x_i \notin L_l \end{cases} \tag{5}$$

$$\varepsilon = \frac{errorlNum}{textNum} \tag{6}$$

其中：x_i 表示测试样本，L_l 表示在类别 l 下测试样本的分类器权重。*errorlNum* 为类别 l 下的分类错误率。A 如式(4)所示。

因此，通过上述数据的分析，得出了类别加权分类器权重方法能更好的表示出分类器权重。将类别加权分类器权重方法融入到多类型分类器方法模型中，得出了改进的多类型分类器模型，模型如图 6 所示。

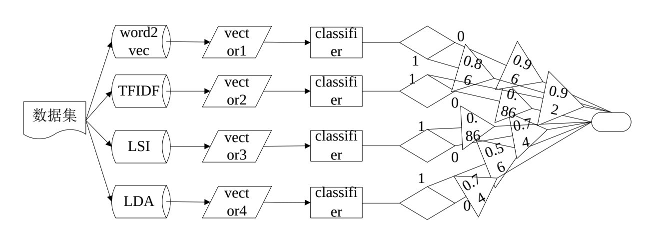


图 6　多类型二元分类器融合示意图

1.4　多类型分类器的算法步骤

输入：样本训练集 *x_train*，样本测试集 *x_test*，样本训练集标签 *y_train*，样本测试集标签 *y_test*，分类器数目 *classNum*

输出：预测结果矩阵 *predicted*

```
1)  calculate word2vec for x_train as class1
2)  calculate tfidf for x_train as class2
3)  calculate lsi for x_train as class3
4)  calculate lda for x_train as class4
5)  train classifier according to x_train
6)   for i∈{1,2,…,len(x_train)} do
7)     calculate errorword2vecNum、errortfidfnum、errorlsiNum、errorldaNum for
each class according to y_train
8)  end
9)  for i∈{1,2,…,len(x_test)} do
10)    calculate w_ij for errorword2vecNum、errortfidfnum、errorlsiNum、errorldaNum
according to equation (4)、(5)and(6)
11)  for i∈{1,2,…,len(x_test)} do
12)    for j∈{1,2,…,classNum} do
13)      s[j] += class[j]*w[j][i]
14)    end
15)    predicted[i] = maximum number index for s is class
16)  end
17)  return predicted
```

2　实验与结果分析

实验平台基于 anaconda 平台，编程语言为 python 语言，4 GB 内存和 1 TB 硬盘的电脑上进行实验。

2.1　实验数据

为了验证多类型分类融合方法的性能，本文利用 nltk 中的 movie_reviews 的特定场景的语料库[13]和普通文本分类搜狗语料库以及 20news 语料库进行实验验证[14]。其中：搜狗语料库以及 20news 语料库是最常用的文本分类语料库，可用于测试算法的不同性能。而 20news 的数据是相对平衡的数据集。movie_reviews 是关于电影评论的情感分析语料库，用于情感分析等分类工作，通过使用 movie_reviews 能更好的验证本文算法。数据集的分布如表 1 所示。

表 1　数据集分布

数据集名称	类别名	训练集/个	测试集/个
Movie_reviews	pos	1600	400
	neg	1600	400
20news	atheism	640	160
	med	792	198
	crypt	794	198
	graphics	800	200
搜狗语料库	社会	1460	365
	娱乐	1460	365

2.2　实验分析

该部分的实验分为三个实验来进行的，第一个实验主要验证了算法的有效性，第二个实验主要关注特征维数对本文算法的影响，第三个实验验证类别加权分类器权重计算方法的有效性。分类方法采用 python 语言中 sklearn 库中的 KNeighborsClassifier 方法作为分类方法(k 值为 10)，特征提取方法有 word2vec、LSI、LDA 和 TFIDF 方法，过滤掉小于 30 的特征词，使用样本识别率作为分类的评价标准，并采用 6 折交叉验证方法验证算法的有效性[15]。本文实验将多类型融合方法以及其他下属混合方法进行实验，以验证本文算法的性能。

2.2.1 多类型分类器融合方法的实验对比

该部分使用了 20news 和 movie_reviews 数据集来进行实验的，20news 为相对平衡数据集侧重于多元分类，movie_reviews 为平衡数据集侧重于二元分类和特定场景的分类，特征维数采用的是 300 维特征，实验结果如表 2 所示。

表 2　20news 和 movie_reviews 分类结果

	Movie_reviews(%)	20news(%)
LDA	57.58	92.06
TFIDF	70.13	93.25
LSI	69.46	93.85
word2vec	62.71	79.89
LDA+TFIDF	65.33	93.72
LSI+TFIDF	71.21	94.44
LSI+LDA	65.75	93.32
word2vec+TFIDF	69.58	89.15
word2vec+LDA	61.54	87.37
word2vec+LSI	69.21	88.49
LSI+LDA+TFIDF	70.75	95.44
word2vec+TFIDF+LDA	69.38	95.11
word2vec+LSI+TFIDF	70.75	94.58
word2vec+LSI+LDA	69.38	95.30
word2vec+LSI+TFIDF+LDA	73.34	96.49
平均值	67.74	92.16
本文算法的最小识别率	2.13	1.06

通过表 2 可知，多元分类 20news 的平均识别率为 92.16% 比 movie_reviews 的 67.74%高。这是因为 movie_reviews 是专业情感分析数据集而 20news 为文本分类数据集，应用的领域不同。本文的多类型分类器方法以及多类型分类器下属方法相比，movie_reviews 取得了良好的效果，最低提升了 2.13%，20news 最低提升了 1.06%。这是因为 20news 是相对平衡的数据集，所以分类器的识别率倾向于多样本类别。

2.2.2 特征维数对融合分类器的实验影响

该部分使用 movie_reviews 来验证不同维数(100、300、500 和 700)对分类性能的影响，实验结果如表 3 所示。

从表 3 可以看出，movie_reviews 数据集的特征维数为 300 时，多类型分类器的性能最好；随着特征维数的不断增加，分类器的平均识别率有所下降，这是因为特征数量不断增加，使得表示文档的特征向量空间中会出现大量的 0，导致文本特征

空间的稀疏而影响了分类器的效果。

表 3　movie_reviews 不同特征维数

	100	300	500	700
LDA	57.75	57.58	57.50	57.58
TFIDF	68.70	70.13	68.75	69.00
LSI	68.95	69.46	68.75	69.13
word2vec	61.65	62.71	61.55	62.04
LDA+TFIDF	64.50	65.33	64.45	64.00
LSI+TFIDF	71.00	71.21	69.85	69.92
LSI+LDA	64.90	65.75	64.30	64.54
word2vec+TFIDF	68.40	69.58	68.30	67.42
word2vec+LDA	60.55	61.54	60.55	59.96
word2vec+LSI	69.35	69.21	68.60	67.42
LSI+LDA+TFIDF	70.20	70.75	69.30	69.63
word2vec+TFIDF+LDA	68.55	69.38	68.00	68.68
word2vec+LSI+TFIDF	70.45	70.75	69.85	69.58
word2vec+LSI+LDA	68.60	68.75	68.05	67.92
word2vec+LSI+TFIDF+LDA	72.15	73.00	70.60	70.75
平均值	67.05	67.65	66.56	66.48
本文算法的最小识别率	1.15	2.13	0.75	0.83

2.2.3 类别加权分类权重的实验结果比较

该部分使用搜狗数据集来验证类别加权分类器权重的有效性，实验结果如表 4 所示。

表 4　搜狗语料库结果

	分类性能加权	类别加权
LDA	95.11	96.21
TFIDF	95.07	95.57
LSI	95.80	95.98
word2vec	89.59	91.74
LDA+TFIDF	89.59	91.74
LSI+TFIDF	94.75	95.57
LSI+LDA	89.59	91.78
word2vec+TFIDF	95.98	95.48
word2vec+LDA	89.73	91.78
word2vec+LSI	95.80	95.57
LSI+LDA+TFIDF	94.75	94.16
word2vec+TFIDF+LDA	89.59	95.11
word2vec+LSI+TFIDF	96.07	96.94
word2vec+LSI+LDA	96.07	96.85
word2vec+LSI+TFIDF+LDA	97.40	98.22
平均值	93.66	94.85
本文最小识别率	1.32	1.28

从表 4 中可以看出，类别加权相对分类性能加权方法具有一定的效果。本章算法平均正确率上比分类性能加权方法高 1.19%，分类器融合方法比下属方法也高出 0.82%，在单独特征提取方法中分类性能均得到了提高了，只有在 word2vec+TFIDF 和 word2vec+LSI 以及 LSI+LDA+TFIDF 方法效果没有提高，这是因为语料库的数据规模使得 word2vec 方法的词向量得分类效果不是很好，同时 LDA、LSI 以及 word2vec 所取得的识别率几乎相同，使得 word2vec+LSI+TFIDF 的融合方法和 word2vec+LSI+TFIDF+LDA 的融合方法效果相差无几，从而影响了融合分类器整体的分类性能。

3　结束语

本文针对单一的分类器以及单一的特征提取方法没有很好的扩展性问题，提出了多类型分类器融合的文本分类方法。本文以四个不同类型的特征提取方法融合起来组成了多类型的文本分类方法。并对分类器权重没有考虑类别信息的问题，提出了类别加权的分类器权重计算方法。通过二元分类和多元分类实验来验证本文算法的有效性。下一步工作是将该方法如何并行计算，减少模型的计算时间。

参考文献：

[1] 何力，丁兆云，贾焰，等. 大规模层次分类中的候选类别搜索 [J]. 计算机学报, 2014, 37 (1): 41-49.

[2] 李荣陆，王建会，陈晓云，等. 使用最大熵模型进行中文文本分类 [J]. 计算机研究与发展, 2005, 42 (1): 94-101.

[3] 黄文明，莫阳. 基于文本加权 KNN 算法的中文垃圾短信过滤 [J]. 计算机工程, 2017, 43 (3): 193-199.

[4] Goudjil M, Koudil M, Bedda M, et al. A novel active learning method using SVM for text classification [J]. International Journal of Automation and Computing, 2016: 1-9.

[5] Jiang L, Li C, Wang S, et al. Deep feature weighting for naive Bayes and its application to text classification [J]. Engineering Applications of Artificial Intelligence, 2016, 52 (C): 26-39.

[6] Sangodiah A, Ahmad R, Wan F W A. A review in feature extraction approach in question classification using Support Vector Machine [C]// Proc of IEEE International Conference on Control System, Computing and Engineering. 2015.

[7] Chen S, Jin Z. Weibo topic detection based on improved TF-IDF algorithm [J]. Science and Technology Review, 2016, 34 (2): 282-286.

[8] Wang Z, Ma L, Zhang Y. A hybrid document feature extraction method using latent dirichlet allocation and word2vec [C]// Proc of IEEE International Conference on Data Science in Cyberspace. 2016.

[9] Santosh D T, Babu K S, Prasad S D V, et al. Opinion mining of online product reviews from traditional lda topic clusters using feature ontology tree and Sentiwordnet [C]// Proc of International Conference on Social Computing and Social Media. 2016: 34-44.

[10] Uysal A K, Gunal S. Text classification using genetic algorithm oriented latent semantic features [J]. Expert Systems with Applications, 2014, 41 (13):

5938-5947.

[11] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [C]// Proc of International Conference on Learning Representations. 2013.

[12] 王晓丹，李睿，薛爱军，等. 基于熵的自适应加权投票 HRRP 融合识别方法 [J]. 系统工程与电子技术, 2017, 39 (4): 707-713.

[13] Jongeling R, Sarkar P, Datta S, et al. On negative results when using sentiment analysis tools for software engineering research [J]. Empirical Software Engineering, . 2017, 22 (5): 2543-2584.

[14] 魏勇，胡丹露，郝晨光，等. 基于分类关键词词频模型的地缘政治主题爬虫设计 [J]. 计算机工程, 2016, 42 (2): 45-50.

[15] 段宏湘，张秋余，张墨逸. 基于归一化互信息的 FCBF 特征选择算法 [J]. 华中科技大学学报：自然科学版, 2017, 45 (1): 52-56.